DOCUMENT RESUME

ED 413 751 FL 024 772

AUTHOR Paskaleva, Elena
TITLE European Language Resources and the Treasury of the Computerised Russian Language Fund (a Small Project Provoking a Discussion on a Big Issue).
PUB DATE 1995-00-00
NOTE 6p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computational Linguistics; Computer Software; *Computer Software Development; *Dictionaries; Foreign Countries; Information Sources; *Language Research; Linguistic Theory; *Russian; Structural Analysis (Linguistics); Uncommonly Taught Languages
IDENTIFIERS Europe; Europe (East); *Language Corpora

ABSTRACT

This paper discusses inclusion of the Russian language in European language resource development. It is suggested that European initiatives to develop language resources can not afford to ignore Russian, and that there is much work to do to resolve differences in resources on the two sides of the former Iron Curtain. A 1995 project, undertaken by organizations in Russia, Germany, and Bulgaria, was designed to make use of the 10,000 vocabulary entries from a Russian dictionary in two machine translation systems dealing with syntactic information and to provide software to accelerate conversion of data into an appropriate computer format. The paper provides a brief outline of this project. Finally, the status and characteristics of current Russian archival material are described. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 751

European Language Resources and the Treasury of the Computerised Russian Language Fund (a Small Project Provoking a Discussion on a Big Issue)

Elena Paskaleva

Linguistic Modelling Laboratory
Bulgarian Academy of Sciences
Sofia 1113, 25a, acad.G.Bontchev str.
Tel.: +359 2713 38 41
Fax: +359 2 70 72 73
E-mail: hellen@bgcict.acad.bg.

BEST COPY AVAILABLE

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Norbert
Volz
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL024772

In this presentation, I would like to draw your attention – through a small project of Central European University Research Support Scheme (CEU RSS) – to a larger issue dealing with the treasury of the linguistic resources of the Russian language.

1. Is Russian a European Language?

If a European initiative aspires to cover the treasure of European linguistic resources, it cannot afford to ignore the Russian language. Moreover, many European joint research projects include a number of Slavonic languages, and the only one that has international status is being left out. We all realise that the issue is rather of organisational than of political nature, and so we must give some serious thought to overcoming the problem. It was no accident that the earliest multi-language NLP applications turned towards the Russian language (such as the systems for machine translation in the early 60s, which started with translating "Pravda" and "Izvestija" into English).

2. The Relations between the Russian and the European CL Achievements from a Historical Perspective

During the long time we spent on both sides of the Iron Curtain, the flow of information between Russian and Western linguistic schools was seriously hindered. Obstacles existed in both directions:

a) The West was not thoroughly acquainted with the notable achievements of formal and structural models such as the model Meaning-Text. This is why, some of what is being recently introduced into wideknown models such as HPSG was developed in the above mentioned Russian model twenty years ago. The lack of knowledge of Russian linguistic achievements was more characteristic of American than of European linguistic thought. Tenier's structural syntax was the basis for Russian structural linguistics, and the developed syntactic models agreed with the achievements of the Prague school of linguistics.

b) The broken connection in the West-East direction found expression in the formalisms and in the hardware and software products used by Russian computational linguistics.

In the triad where the achievements of computational linguistics are situated, i.e., a) linguistic structures, b) formal description, and c) computational models, the Russian school was the most stable one in a), but lagged behind

in c) for reasons beyond science. The Western school had arranged its priorities in the reverse order.

Of the Eastern European countries with long traditions in computational linguistics, two are the ones which established relationships with Russian structural and formal linguistics (in its best): Czechoslovakia and Bulgaria. The former primarily because of its leading role in Slavonic language studies and the proximity of the linguistic schools. The latter because of the lexical similarity, the well established traditions in Slavonic language studies and the historically determined lack of Russophobe disposition.

The deficiency of computer resources did not prevent scientists of the Russian Language Institute from the beginning – 15 years ago – to develop a Computerised Russian Language Fund (CRLF) using sofware and hardware lagging a generation behind the Western one. A more careful look at the present CRLF archive (distributed in hyper text form, see below) convinces us that the accumulation of Russian language resources has a long way to go, which is also evident from the wide range of text processing tools used in this archive.

In the international links, however, the bilateral relations between the Russian Academies of Science and the East European countries were cancelled for financial reasons. The common projects were cancelled right at the moment when the unification of software products and the presentation of linguistic knowledge had begun.

The revival of contacts started not long ago following schemes rather different from the old ones. In the trans-European scientific cooperation, Russia was separated from the East European countries as a participant in a the special program INTAS. Presently, the funding of joint projects with participants from the West, Eastern Europe, and Russia is accidental within the frameworks of wider initiatives such as the Open Society Fund.

3. A Brief Outline of the Small Research Project

In 1995, CEU RSS (Central European University – Research Support Scheme) sponsored a small project with three participants: CRLF, Moscow; GMS (Gesellschaft fur multilinguale Systeme), Berlin, and LLM (Linguistic Modelling Laboratory, Bulgarian Academy of Sciences), Sofia. These limited resources have been granted for a project with a quite imposing title:

Application of the Data of the Computerised Russian Language Fund to the NLP Systems.

The word "apply" is a three arguments' predicate (subject- object- object). In the title above, the last two arguments need explanation. What is meant here by "data of CRLF" is the use of 10 000 vocabulary entries from Ozhegov's Dictionary of the Russian language. This dictionary is recognized by the contemporary Slavonic lexicographers as one of the best modern uni-lingual, medium-sized Russian dictionaries. "NLP systems" means MT systems like METAL and EUROTRA which execute machine translation.

The linguist researcher is attracted in this project by the requirement for unifying the parameters of the linguistic knowledge represented with the appropriate depth and scope in two real products. METAL and EUROTRA are machine translation systems dealing with syntactic information. A considerable part of the lexical entry in Ozhegov's dictionary is dedicated to the syntactic valency of the lexical item. The adjustment of the Russian data to the above systems becomes easier due also to the fact that they function as DB based on the special software UNILEX created in CRLF. These tools facilitate the transition between the two types of products with the help of an intermediate representation of, so to say, a generalized dictionary (with a temporary title ADALEX).

The Laboratory for Linguistic Modelling will participate with software products accelerating the conversion of data into the appropriate format. The design of the general ADALEX dictionary will include a permanent text support for the collection of data about the subcategorization. This requires the development of tools for textual support. The project is by no means a project for the joint work of institutions – it is simply the joint work of individual scientists towards a particular task.

The ideology of this project is in harmony with the developed trans-European projects for unification of the dictionary formats (like EAGLES and GENELEX). Thus, this project is a small model of the future real inclusion of the wealth of Russian language resources in the Europe-wide programs for linguistic technologies.

4. A Glimpse of the Actual Resources of CRLF

The modest dimensions of the resources of this project are evident when compared to the current resources of CRLF kindly presented to me by my Russian partners in the hypertext format [see(1) and (2)].

Unfortunately, the available description of the archive is in Russian, but its presentation here proves that the text- processing software products currently used in Russia are convertible in Europe. This was not the case 20 years ago.

The variety of linguistic resources in the archive is realy impressive.

Textual resources are represented in three ways: as plain text (texts in DOS and WINDOWS formats), as marked text (with manual SGML-like annotation), and as DBF files (created and supported by a special DB tool - UNILEX).

Data is stored on magnetic tapes and diskettes.

Most of the files are archived.

The collection of literary texts is striking - from Tolstoj to Brodsky.

The most precious stone in this treasury is the collection of dictionaries: most of them are in DB format. The collection includes monolingual, orthographical (general and special), syntactic, grammatical, and morpheme dictionaries of the Russian language.

The entering of the texts (in the course of 15 years) ranges from manual work to OCR processing.

In order not to seem like advertising, the above statements are accompanied by diskettes containing the description of this archive which are placed at the disposal of all participants in the workshop with the consent of CRLF.

References

L. Kolodjazhnaja. The archive of linguistic resources of the CLRF (in Russian, in electronic form).

The Russian Language Computerized Fund. Bulletin No 3, Moscow 1995 (in Russian, in electronic form).